DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4703

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

REVIEW PAPER

Available Online at www.ijarcs.info

INTRUSION DETECTION SYSTEMS: A REVIEW

D. Ashok Kumar
Associate Professor, Department of Computer Science
Government Arts College, Thuvakudimalai,
Tiruchirappalli, India

S. R. Venugopalan
Scientist, Information and Computing technologies
Aeronautical Development Agency (Ministry of Defence)
Bangalore, India

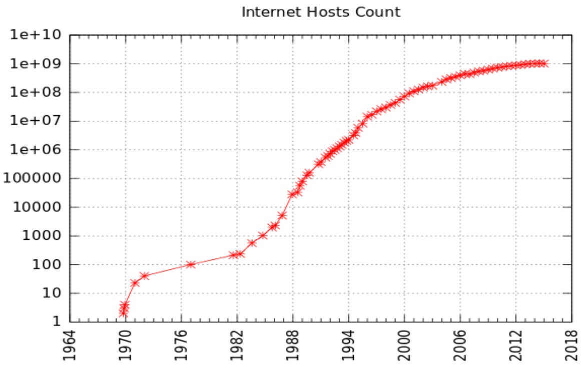
Abstract: Given the exponential growth of Internet and increased availability of bandwidth, Intrusion Detection has become the critical component of Information Security and the importance of secure networks has tremendously increased. Though the concept of Intrusion Detection was introduced by James Anderson J. P. in the year 1980, it has gained lots of importance in the recent years because of the recent attacks on the IT infrastructure. The main objective of this study is to examine the existing literature on various approaches for Intrusion Detection in particular Anomaly Detection, to examine their conceptual foundations, to taxonomize the Intrusion Detection System (IDS) and to develop a morphological framework for IDS for easy understanding. In this study a detailed survey of IDS from the initial days, the development of IDS, architectures, components are presented.

Keywords: Network traffic, Information security; Intrusion detection; Attack, Network anomaly detection, Taxonomy

1. INTRODUCTION

Threats in the Internet are posing higher risk on Security of Information. Intrusion Detection is the process of monitoring and analyzing network traffic and events in the system to detect any vulnerabilities and attacks. Now Intrusion Detection has become the priority and an important task of Information Security administrators. A system deployed in a network is vulnerable to various attacks and needs to be protected against attacks [2]. Intrusion Detection Systems (IDS) play a vital role in protecting organization's security. Intrusion is a deliberate unauthorized, illegal attempt to access, manipulate or taking possession of an Information System /Network to render them unreliable or unusable. Intrusion Detection is the process of finding important events occurring in a system and analyzing them for possible presence of Intrusion. Intrusion Detection Systems are implemented using Hardware and/or Software. The aim of an Intrusion Detection System (IDS) is to protect the system from unauthorized access and Intrusion Detection is the process of identifying various events occurring in a system/network and analyzing them for possible presence of Intrusion and responding to the malicious activities. Now most of the attacks/intrusions are network based and the network needs to be protected. Researchers have used various approaches such as data mining, soft computing, Machine Learning, Statistical Techniques, Bayesian Techniques, Artificial Neural Networks and Evolutionary Computing etc. for Network Anomaly Detection have achieved performance improvements. The combination of various approached have been used by researchers for further improving the performance. The growth of Internet can be well-understood by the following Fig. 1. In 2017 there are approximately around 1E+09 hosts across the globe and this keeps growing.

The growth of Internet has brought great benefits to the society at the same time the growing attacks on the IT Infrastructure are becoming an increasingly serious issue and needs to be addressed. Along with the growth of Internet attacks are also growing in parallel.



Source:https://en.wikipedia.org/wiki/Global_Internet_usage accessed on 1st Aug 2017

Fig. 1: Growth of Internet in terms of Host Count

In earlier days, the attacker should have a good knowledge about the target infrastructure and knowledge on the Network, Operating Systems & Applications. Whereas today there are lots of open tools available in the Internet which can trigger automated attacks. From the following Fig. 2, it can be seen that the sophistication of attacks are increasing while the need for need of Intruder knowledge is reducing. Attacks range from simple viruses, worms to malwares, Denial of Service (DOS), Network Attacks and Ransomware Attacks. There are several type attacks that do not attack computers but rather attacks on the networks such as flooding.

Fig: 2. Attack sophistication Vs Intruder Knowledge

The recent "WannaCry worm travelled automatically between computers without user interaction.

WannaCry Attack: In May 2017, the WannaCry Ransomware spread through the Internet, using an exploit vector named EternalBlue. The ransomware attack infected more than 230,000 computers in over 150 countries using 20 different languages to demand money from users using Bitcoin cryptocurrency. WannaCry demanded US$300 per computer.

Petya Attack: Petya worm spread during April 2016, this malware infected the master boot record of the computer by encrypting the file tables of NTFS file system. Once infected on the next boot expects a ransom is paid. Again in the month of June 2017, modified version of Petya using EternalBlue exploit and this was aimed to create disruption rather to generate profit.

The massive cyber-attack in history, the WannaCry malware majorly targeted healthcare and government infrastructure in the western countries. India too, felt the heat of WannaCry, however, the level of damage remained substantially less compared to other countries.

Various researches have studied and surveyed Intrusion Detection System at various levels. The study of the existing literature reveals that there is a need for an up-to-date and detailed survey of Intrusion Detection Systems.

2. NEED FOR INTRUSION DETECTION SYSTEMS

An Organization's Information security is comprehensively enabled by IDS that comprise a set of integrated software and hardware that are blended into the organizational Information/Data policies and practices concerning the Security.

When a system is deployed in the Internet, it is vulnerable to various attacks and the system needs to be protected against various attacks. Any activity that attempt to compromise the confidentiality or integrity, or availability of a resource is termed as Intrusion [3]. Increased use of computer networks, internet and online transactions pose higher risk of intrusions and protecting the information from the hackers/intruders is a new area in computers and network security. Intrusion Detection has become an inevitable area for commercial applications and academic research. The major factors which affect intrusion detection are the system's detection rate and time required to detect intrusions. Many researchers have focused in this area and have used Data Mining & Machine Learning techniques for detecting the intrusions. A computer system or network is said to be reliable if confidentiality, integrity and availability is a part of its security requirements [4]).

Confidentiality requires that information can be accessible only to those authorized for it

Integrity requires that information remain unchanged without any modification by malicious attempts;

Availability means the computer system and its resources are always available to authorized users when they need it.

Intrusion Detection System (IDS) plays a vital role in preserving the data integrity, confidentiality and system availability from attacks.

The security solutions like firewall are not designed to handle network or application layer attacks such as DoS, DDos, Worms, Viruses and Trojans. The growth/wide-spread of Internet and the prevalent threats are the reasons for deploying IDS. IDS operate behind a firewall looking for malicious activity. A firewall is a network security system that monitors/prevents a specific type of information from moving between the untrusted network outside and the trusted network inside. Firewall is a gatekeeper computer between the Internet and a private network and protects the private network by filtering traffic to and from the Internet based on defined policies (rules). The firewall may be a separate computer system, a service running on an existing router or server, a separate network containing a number of supporting devices. Firewalls are often categorized as Packet-filtering firewalls and Application-level firewalls.

Firewalls are setup to stop unnecessary network traffic into or out of any network. The network traffic approaching a firewall is either allowed or stopped according to the configured rule. Whereas an intrusion detection system gathers and analyses information from various areas within a computer or a network to identify possible security breaches.

3. ISSUES, LIMITATIONS OF THE PRESENT DAY IDS AND ITS CHALLENGES

With the ever increasing deployment of 10G/1G networks, the traditional IDSs have not scaled accordingly. The availability of higher bandwidth and sophisticated hardware and software, the need to detect intrusions in real-time and the adaptation of the detection algorithm to the ever changing traffic pattern is a big challenge. The increasing size and complexity of the Internet along with variety end hosts systems make it more prone to vulnerabilities. With present Hardware, it is becoming difficult to detect intrusions in real-time.

Data overload: Number of devices which access Internet has increased tremendously. It is extremely important that how much data and IDS can efficiently handle. In the present days, the data transfer/access has increased because of higher bandwidth and easy access to information with Mobiles and hand-held devices via 3G and 4G networks.

Encrypted traffic: The process of Intrusion Detection is made more difficult with the use of encrypted traffic

False positives: A false positive occurs when normal traffic is mistakenly classified as malicious and treated accordingly.

False negatives: In this case, IDS does not generate an alert when an intrusion has actually taking place and malicious traffic is classified as normal.

There are lots of IDSs available both commercially and in public domain. These IDSs use different approaches to detect intrusions and each of these have shown distinct preferences over certain classes of attacks. The analysis of these IDSs shows that there are problems which are to be solved before the development of IDS which is reliable and can detect wide range of intrusions. The following are some of challenges before us in the development of IDS.

- ✓ With the increased speed and bandwidth, capturing all network packets and processing them in real-time is a challenge.
- ✓ With the ever increasing deployment and usage of 1G/10G networks, traditional network anomaly detection

based intrusion detection systems have not scaled accordingly

- ✓ Anomaly detection systems suffer from high false alarm. Reduced number of false alarm defines the usability of Network Anomaly Detection Systems
- ✓ Designing a generic Anomaly Detection System which can work in all environments is a challenge because each environment and the security requirements are unique.
- ✓ The ever changing network demands the ADS should be adaptive. Developing adaptive ADS is a challenge because the intruders change their strategy and adapt to the ever changing networked environment.
- ✓ As zero day attacks are becoming prominent and new vulnerabilities and exploits are discovered every day, the ADS should be adaptive to unknown attacks.
- ✓ Nowadays, distributed attacks are becoming prominent and attacks can compromise thousands of system within no time and the ADS should be capable of handling mass attacks.
- ✓ The non-availability of recent ground truth dataset with all the recent attacks that captures the real networks makes the ADS design a challenging issue.
- ✓ Researches use various approaches to design ADS. But the approach is tested with a dataset and this approach may not work with other data sets.
- ✓ The main challenge is to define a model for normal and attack traffic which can cater to the changes in the network behaviours over time

A Simple Google Search of ‘network anomaly detection’ has showed 20, 30,000 items and “scholar. Google” showed 323,000 items on July 22, 2017

Security is the quality or state of being secure i.e. free from danger. Security can be classified into Physical Security, Personnel Security, Operations Security, Communications Security, Computer Security, Network Security and Information security. The concept of security has undergone several changes, initially the data processing was centralized and the security is limited to the centralized system. With the widespread of distributed data processing systems and networks, the landscape of security has changed. It is required to provide security to the centralized system, end hosts and as well as the data which is in transit/transmission in the Network.

4. IDS AND ITS EARLY HISTORY

According to Kruegel et al. (2004), “intrusion detection is the process of identifying and responding to malicious activities targeted at computing and network resources” [5]. An intrusion attempt, also named as attack refers to a sequence of actions by means of which an intruder attempts to gain control of a system [6]. The aim of an Intrusion Detection System (IDS) is therefore to discriminate intrusion attempts and intrusion preparation from normal system usage.

Network security measures were required to protect data during transmission. Network security involves protecting a network from unauthorized access and risks. It is the duty of network administrators/Information Security specialists to adopt preventive measures to protect their networks from potential security threats. To detect unauthorized activities in the network or on individual machines, organizations implement Intrusion Detection Systems (IDS). Intrusion detection has been studied for almost 20 years. Intrusion Detection Systems monitor malicious/unauthorized activities in a network, log information about such activities/send alarms, take steps to stop them/drop packets, and finally report them.

IDS was originally introduced by Anderson in his “Computer Security Threat monitoring and Surveillance” [1] and later the same was formalized by Denning in 1987 [7]. Denning’s paper describes a model for real-time intrusion detection expert system capable of detecting attacks. Between 1984 and 1988, the prototype Intrusion Detection Expert System (IDES) was developed which was first used to monitor activities of users and this was capable of providing real-time detection of security violations on single-target host systems. This IDES was initially a rule-based expert system trained to detect known malicious activity. IDES was a critical first step towards the development of real-time dual-analysis (signature analysis and anomaly-detection) intrusion-detection. This same system has been refined and enhanced to what is known today as the Next-Generation Intrusion Detection Expert System (NIDES). In 1988, “Haystack” was the first IDS to use patterns and statistical analysis for detecting malicious activities but lacked real-time analysis [8]. An X/Motif-based graphical user interface facility was added to NIDES to provide location-independent configuration and monitoring of NIDES operation and it greatly increased usability.

The significant milestone in Intrusion detection was the development of Network System Monitor (NSM) at the University of California at Davis Lawrence Livermore Laboratories in 1990, for monitoring network traffic. Later this was developed into “Distributed Intrusion Detection Systems (DIDS)” in 1992. US Government funded projects like Discovery, Haystack, Multics Intrusion Detection and Alerting System (MIDAS), Network Audit Director and Intrusion Reporter (NADIR) to develop intrusion detection systems [9]. In 1991, Network Audit Director and Intrusion Reporter (NADIR) were developed to monitor user activities on Integrated Computing Network (ICN). NADIR combines the rule-based analysis and statistical profiling. USTAT IDS was developed for UNIX in 1993 [10]. Use of autonomous agents was suggested by the authors of [11] in order to improve the scalability, maintainability, efficiency and fault tolerance of an Intrusion Detection System.

Next Generation IDES (NIDES) was developed in 1995 which succeeded IDES project. It is an anomaly detection system using statistics and has signature based component also [12]. In the Mid 90’s, SAIC developed “Computer Misuse Detection System” (CMDS), a host based IDS. US Air Force’s Cryptographic support centre developed “Automated Security Incident Measurement” (ASIM), which addressed the issues like scalability and portability. “Stalker” based on DIDS became the first commercially available IDS and influenced the growth and trends of future IDS.

A stand-alone system named *Bro* was introduced for detecting real-time network intrusions [13]. Bro monitors the network passively and it also provides a real-time notification of ongoing or attempted attacks. The IDS market began around 1997 and "Real Secure" network intrusion detection was developed by ISS. In 1998, Cisco purchased Wheel Group to provide security solution to their customers. From there, the commercial IDS world expanded its market-base. SNORT an Open-Source Network IDS was launched in 1998 [51] which gained much popularity. In year 1999 Okena Systems worked out the first Intrusion Prevention System (IPS) under the name "Storm Watch". IDSs started sharing information to discover attacks involving multiple locations. A framework was presented on how the IDSs can share information to discover attacks [14]. Fuzzy Intrusion Recognition Engine (FIRE) an anomaly based IDS that uses fuzzy logic to detect whether the activity is malicious or not was proposed and this uses simple data mining techniques [15].

Neural networks for intrusion detection were first introduced as an alternative to statistical techniques in the IDES intrusion detection expert system [16]. Ghosh et al. found that a "well trained, pure feed forward, back propagation neural network" performed comparably to a basic signature matching system [17]. Various neural networks can be used for anomaly based IDS like Multi layered Perceptrons, Radial Basis Function-Based, Hopfield Networks etc.

Bayesian approaches are very popular in anomaly detection. The results of Bayesian techniques are similar to those of Threshold based systems but the computational resource consumption is high [18].

Jiong Zhang and Mohammed Zulkernine in their paper proposed an unsupervised framework for anomaly detection based on outlier detection techniques in random forests algorithm [19]. A frame work was built on the patterns on network services and the same was used to detect attacks using modified outlier detection algorithm, reducing the calculation complexity.

5. ATTAKS AND ITS TAXONOMY

An attack is a set of operations that puts a system under security risk. Attacks can be classified into eight main categories [20].

Physical attacks: These attacks involve damaging the computers and network hardware.

Infection: Some unwanted programmes are installed on the target system that may corrupt the system or utilize the system resources. Eg. Viruses, Worms and Malwares.

Exploding: This category of attacks seeks to explode or overflow the target system with bugs. Eg. Buffer Overflow

Probe: This type of attacks collect the information about the target system. Port Scanning, Sniffing.

Cheat: Fake identities are used to get into the system. Eg. Session Hijacking, XSS, IP/MAC Spoofing

Traverse: This category of attacks uses all possible ways to match the system credentials to get into the system. Eg. Brute Force, Dictionary Attacks.

Concurrency: This system of attacks compromise the system availability by sending mass requests that the system cannot handle. E.g. Flooding, DoS, DDoS,

Others: These are attacks which uses the known vulnerability/weakness to compromise the system. This does not require any professional skills. The target is the systems which are not configured properly

Further attacks can be classified into passive and active. Passive attacks are launched to gather information and monitor network traffic. By using this information active attacks can be initiated. Port Scans, Sniffing are examples of passive attacks. Active attacks are classified into four categories by Defence Advanced Projects Agency (DARPA) namely

DOS: Denial of Service Attacks are designed to disturb the host and availability of host or service is compromised.

Probe: These attacks scan the computer or network to gather useful information about the hosts, valid IP address, active ports and OS etc. [21]. The information gathered are used to launch attacks.

R2L: In this type of attacks the user who does not have account in the system gains local access.

U2R: In this type of attacks the user who has local account gains privileges of super user account.

The following Table 1 lists the attacks based on DARPA classification.

Table 1: Attacks in DARPA grouped into various categories

Attack Taxonomy	***Attacks***
DOS	*Apache, back, land, mailbomb, netptune, pod, processtable, smurf, teardrop, udpstrom*
Probe	*Ipsweep, mscan, nmap, saint, satan, portsweep*
U2R	*Bufferoverflow, loadmodule, perl, ps, rootkit, sqlattack, xterm*
R2L	*ftp_write, guess_password, httptunnel, imap, multihop, named, phf, sendmail. snmpgetattack, snmpguesss, warzmaster, worm, xclock, xsnoop*

The following are the some of the commonly used attacks and these attacks are discussed brief.

Malware is short for malicious software and used as a single term to refer virus, spy ware, worm etc. Malware is designed to cause damage to a standalone computer or a networked System. Malware is a term used to refer a program which is designed to damage the computer; it may be a virus, worm or trojan.

Viruses: A computer virus is a program, script, or macro designed to cause damage, steal personal information, modify data, send e-mail, display messages, or some combination of these actions.

Worms: A worm is a destructive self-replicating program containing code capable of gaining access to computers or networks. Once it gains access to computer or network, the worm causes harm by deleting, modifying, distributing, or otherwise manipulating data.

Spywares: A term used to describe a software program that has been designed to secretly gather information about a user's activity. Spyware programs are often used to track users' habits to better target them with advertisements. Spyware is usually installed onto a user's machine without their knowledge when a link is followed (intentionally or unintentionally) which redirects the user to a malicious website.

Adware: Alternatively referred to as malware, sneak ware, or spyware, adware is a program installed without a user's consent or knowledge during the install of another program. Much like spyware, adware tracks individuals Internet activities activities and habits to help companies for advertising more efficiently.

Trojans: A trojan horse is a program that appears to be something safe, but in is performing tasks such as giving access to your computer or sending personal information to other computers. Trojan horses are one of the most common methods a criminal uses to infect the computer and collect personal information from the computer.

Web Trojans: Web Trojans are malicious programs that pop up over login screens to collect credentials. The user believes that he or she is entering information on a website, while in fact the information is being entered locally, and then transmitted to the attacker for misuse.

Ransomware attacks: Ransomware is a type of malware that blocks access to a computer or its data and demands money to release it.

CoolWebSerarch: CoolWebSearch is malicious software that once it is executed has the capability of replicating itself and infects other files and programs. This type of malware can steal hard disk space and memory and slows down or completely halts the PC. It can also corrupt or delete data, erase the entire hard drive, steal personal information, hijack screen and spam contacts to spread it to other users. Usually, a Virus is received as an attachment on an email or instant message.

Spam: Alternatively referred to as UCE (Unsolicited Commercial Email) and bulk e-mail, Spam is slang commonly used to describe junk e-mail on the Internet. Spam is e-mail sent to thousands and sometimes millions of people without prior approval, promoting a particular product, service or a scam to get other people's money.

Phishing: Phishing is a term used to describe a malicious individual or group of individuals who scam users. They do so by sending e-mails or creating web pages that are designed to collect an individual's online bank, credit card, or other login information. Because these e-mails and web pages look like legitimate companies users trust them and enter their personal information.

Denial of Service (DoS): DoS attack is a method of attacking a networked computer by sending it an abnormally high number of requests, causing its network to slow down or fail. Since a single individual cannot generate enough traffic for a DoS attack, these attacks are usually run from multiple computers infected by worms or zombie computers for a DDoS.

Distributed DOS (DDoS): DDoS attacks disturb the normal function of a specific website. That means the attack isn't random, such as a launched virus that's aimed at everyone and anyone but no one in particular. A DDoS is planned and coordinated, and the goal is to make an entire website unavailable to its regular visitors or customers.

Buffer/Stack Overflow: A buffer overflow is an exploit that takes advantage of a program that is waiting on a user's input. There are two main types of buffer overflow attacks: stack based and heap based. Heap-based attacks flood the memory space reserved for a program, but the difficulty involved with performing such an attack makes them rare. Stack-based buffer overflows are by far the most common.

In a stack-based buffer overrun, the program being exploited uses a memory object known as a stack to store user input. Normally, the stack is empty until the program requires user input. At that point, the program writes a return memory address to the stack and then the user's input is placed on top of it. When the stack is processed, the user's input gets sent to the return address specified by the program.

Man in the Middle (MITM): A man-in-the-middle attack is an attack where a user gets between the sender and receiver of information and sniffs any information being sent. In some cases, users may be sending unencrypted data, which means the man-in-the-middle (MITM) can obtain any unencrypted information.

Sniffing: A packet sniffer is a utility that has been used since the original release of Ethernet. Packet sniffing allows individuals to capture data as it is transmitted over a network. This technique is used by network professionals to diagnose network issues, and by malicious users to capture unencrypted data, like usernames and passwords. If this information is captured in transit, a user can gain access to a system or network.

Active Hijacking: Active Session Hijacking means that original user has logged in his account or profile and then attacker steal the cookies to hijack the active session and then disconnect the original user from the server.

DNS Cache Poisoning: DNS cache poisoning, also known as DNS spoofing, is a type of attack that exploits vulnerabilities in the domain name system (DNS) to divert Internet traffic away from legitimate servers and towards fake ones.

DNS Spoofing: DNS Spoofing (sometimes referred to as DNS Cache Poisoning) is an attack whereby a host with no authority is directing a Domain Name Server (DNS) and all of its requests. This basically means that an attacker could redirect all DNS requests, and thus all traffic, to his (or her) machine, manipulating it in a malicious way and possibly stealing data that passes across. This is one of the more dangerous attacks as it is very difficult to detect

DNS amplification: A Domain Name Server (DNS) amplification attack is a popular form of distributed denial of service (DDoS) that relies on the use of publically accessible open DNS servers to overwhelm a victim system with DNS response traffic.

Land attack: In a DoS land (Local Area Network Denial) attack, the attacker sends a TCP SYN spoofed packet where source and destination IPs and ports are set to be identical. When the target machine tries to reply, it enters a loop, repeatedly sending replies to itself which eventually causes the victim machine to crash.

Teardrop attack: A teardrop attack is a denial-of-service (DoS) attack that involves sending fragmented packets to a target machine. Since the machine receiving such packets cannot reassemble them due to a bug in TCP/IP fragmentation reassembly, the packets overlap one another, crashing the target network device. This generally happens on older operating systems such as Windows 3.1x, Windows 95, Windows NT and versions of the Linux kernel prior to 2.1.63.

SYN flood attack: A SYN flood is a form of denial-of-service attack in which an attacker sends a succession of SYN requests to a target's system in an attempt to consume enough server resources to make the system unresponsive to legitimate traffic. This exploits the three way handshake mechanism of TCP. Whenever a client want to connect it sends a SYN packet and the server acknowledges this request by sending SYN-ACK back to the client. This stage the connection is called as half-open connection. The client has to respond with ACK and then the connection is established. Here the client will not send ACK request, so the number of half-open connection increase and the server will not be able to accept new connections.

Smurf attack: The Smurf attack is a distributed denial-of-service attack in which large numbers of Internet Control Message Protocol (ICMP) packets with the intended victim's spoofed source IP are broadcast to a computer network using an IP broadcast address.

Malformed Packet attack: A malformed packet attack occurs when malformed IP packets are sent to a target system, causing the system to work abnormally or break down. With the capability of defending against such attacks, a device can detect and discard malformed packets in real time.

UDP Flood attack: Similar to an ICMP flood, a UDP flood occurs when an attacker sends IP packets containing UDP datagram with the purpose of slowing down the victim to the point that the victim can no longer handle valid connections.

Snork Attack: This attack allows an attacker with minimal resources to cause a remote NT system to consume 100% CPU Usage for an indefinite period of time. It also allows a remote attacker to utilize a very large amount of bandwidth on a remote NT network by inducing vulnerable systems to engage in a continuous bounce of packets between all combinations of systems. This attack is similar to those found in the "Smurf" and "Fraggle" exploits, and is known as the "Snork" attack.

Pharming attacks: Pharming is a cyber-attack intended to redirect a website's traffic to another, fake site. Pharming can be conducted either by changing the hosts file on a victim's computer or by exploitation of a vulnerability in DNS server software.

Identity Theft: Identity theft, also known as identity fraud, is a crime in which an imposter obtains key pieces of personally identifiable information, such as Social Security or driver's license numbers, in order to impersonate someone else.

Bonets: A botnet is a collection of internet-connected devices, which may include PCs, servers, mobile devices and internet of things devices that are infected and controlled by a common type of malware. Users are often unaware of a botnet infecting their system.

IRC Bots: An IRC bot is a set of scripts or an independent program that connects to Internet Relay Chat as a client, and so appears to other IRC users as another user. An IRC bot differs from a regular client in that instead of providing interactive access to IRC for a human user, it performs automated functions.

P2P Bots A peer-to-peer botnet is a decentralized group of malware-compromised machines working together for an attacker’s purpose without their owners’ knowledge.

System Reconfiguration attack: System reconfiguration attacks modify settings on a user’s PC for malicious purposes. For example: URLs in a favorites file might be modified to direct users to look-alike websites: e.g., a bank website URL may be changed from “bankofabc.com”to “bancofabc.com”.

SQL Injection attack: SQL Injection (SQLi) refers to an injection attack wherein an attacker can execute malicious SQL statements (also commonly referred to as a malicious payload) that control a web application’s database server (also commonly referred to as a Relational Database Management System – RDBMS). Since an SQL Injection vulnerability could possibly affect any website or web application that makes use of an SQL-based database, the vulnerability is one of the oldest, most prevalent and most dangerous of web application vulnerabilities.

Cross-Site Scripting (XSS): XSS is a type of computer security vulnerability typically found in web applications. XSS enables attackers to inject client-side scripts into web pages viewed by other users. Across-site scripting vulnerability may be used by attackers to bypass access controls such as the same-origin policy.

Password attack: An attempt to obtain or decrypt a user's password for illegal use. Hackers can use cracking programs, dictionary attacks, and password sniffers in password attacks. Defense against password attacks is rather limited but usually consists of a password policy including a minimum length, unrecognizable words, and frequent changes.

Information gathering/Scanning: The first phase in security assessment is focused on collecting as much information as possible about a target application. Information Gathering is the most critical step of an application security test. The security test should endeavor to test as much of the code base as possible.

Social Engineering attacks: Social engineering is an attack vector that relies heavily on human interaction and often involves tricking people into breaking normal security procedures. It is the process of using social skills to convince people to reveal access information to the attacker. These attacks exploit trust between people and targets large organization. These attacks are often used when attacker cannot find way to penetrate victim's systems using other means. It is very hard preventing without security awareness.

Impersonation attack: An impersonation attack is a form of social engineering in which a bad actor uses electronic communications to "spoof" or impersonate the identity of a trusted colleague. The most common form that impersonation attacks take is in conjunction with phishing emails, which pretend to come from a trusted person or brand in order to trick the recipient into giving cyber criminals money or sensitive information.

6. TAXONOMY OF IDS

IDSs can be classified by various methods such as deployment, detection methods and types etc. This section brief discusses the taxonomy of IDS

A. INTRUSION DETECTION METHODS

Various techniques are in place for intrusion detection which can be broadly classified into Signature/pattern based (or) Misuse (or) Knowledge based detection and Anomaly based (or) behaviour based detection. Both Methods have its own advantages and disadvantages.

Signature based or Misuse Intrusion Detection uses well-defined pattern of the attacks that exploit weaknesses in the system and application software to identify the intrusions. These systems detect intrusions based on a pattern for a malicious activity. It is very useful for detection of known attack patterns, known vulnerabilities on the system. This system compares the network/system activity with the known signatures or other misuse indicators to produce alarms. The rate of missing report is high [22]. Regular updates of signatures are necessary. These systems can also detect intrusion attempts; a partial signature may indicate an intrusion attempt. Examples include Haystatck, Bro, IDES, and Discovery etc.

Anomaly based Intrusion Detection uses the normal usage behaviour patterns to identify the intrusions and is trained using normal behavioral pattern of traffic. Detects malicious activity based on deviations from the normal behaviour are considered as attacks. It can detect unknown intrusions. The rate of missing report is low [22]. These systems build model based over the normal data and then check to see whether the data fits into the model. This system can find unknown attacks. But the False Alarm Rate and accuracy are low when compared with Signature based approach. Anomaly detection system can use either Supervised or Unsupervised learning techniques. Examples are IDES, NIDES, and EMERALD etc. The domain and the nature of anomalies change over time and intruders adapt their network attacks to evade the existing intrusion detection solutions [23].

Supervised Learning: The basic assumption is supervised learning is the availability of training dataset which has labelled instanced for normal as well as for the anomaly classed. Predictive model is normal and anomaly is built using the datasets. The test-case is compared with both the cases to find which class it belongs to. There are two issues in supervised techniques

1) Anomalous instances are few in number in the training data.

2) The quality of labelling process for the training data is another issue

There are lot of artificial techniques to inject anomalies in the training data to get labelled training data.

Unsupervised Learning: These techniques do not require any training data. The basic assumption in this method is the normal instances are more in number than the anomalies. These systems may have large false alarm rate if the above assumption is not true.

Semi supervised Learning: The basic assumption is this technique is that labelled data is only available for normal class. Anything which is away from normal class is assumed to be anomalous.

Protocol Anomaly Detection: Traffic that does not conform to known protocol standards is alerted. This Anomaly detection system checks for any deviation from the normal traffic patterns

Note: There are very few commercial tools which are implemented using this approach, leaving anomaly detection to research systems. It may be noted that the founding paper on IDS (Denning, 1987) advocates this as a requirement for IDS. The following Table 2 lists out the advantages and disadvantages of both the detection methods.

Table 2: Advantages and Disadvantages of Intrusion Detection Methods

Detection Methods	***Advantages***	***Disadvantages***
Signature Based Detection	*1) Is able to detect accurately* *2) Generate much fewer false alarms.*	*1) Cannot detect novel or unknown attacks*
Anomaly Detection	*1) Is able to detect new/unknown attacks based on audit* *2) Less dependent on Operation system specific mechanisms* *3) Can detect abuse of privileges* *4) High False alarm rate*	*1) High false-alarm and limited by training data.* *2) The entire scope of behavior is not usually covered during learning phase.* *3) Behavior change over time causing the system to perform poor.* *4) During learning the system may be undergoing attack which*

		will result in poor results.

Note: There are very few commercial tools which is implemented using this approach, leaving anomaly detection to research systems. It may be noted that the founding paper on IDS, Denning advocates this as a requirement for IDS [7].

B. DEPLOYMENT TECHNIQUES

The positioning of the IDS decides the effectiveness of the Intrusion Detection system.

Host based: IDS is deployed on the end-host, Web servers & database servers and the data from the host is used to detect signs of intrusion. Here the internals of the computing systems such as CPU activity, memory utilization file I/O activity, network activity and Operating System events are analyzed rather than the external interfaces. It monitors the dynamic behavior of the computing system [24]. The internal resource access pattern is monitored to find out the legitimate access. Host based IDS (HIDS) can be thought of as an agent running on the end-host which monitors the events. Examples of HIDS include EMERALD, NFR etc. HIDS instead of monitoring the activity itself, it monitors the audit record of activity. Here the intrusion is detected after the intruder has entered the system. The host resources are used by the HIDS which may affect the actual function of the host.

Network Based: The main function of network based IDS is to monitor and scrutinize the network traffic for any possible intrusions. Intrusions typically occur as anomalous patterns. The Network based IDS (NIDS) reads all the network packets or net flows to find patterns. The data is high-dimensional data with a mix of categorical/discrete and continuous/numerical attributes. Examples include Microsoft Network Monitor, Cisco Secure IDS, and Snort etc. This is easily deployable and there are less performance issues on the monitored host. Since these NIDS runs on a separate system other than the targeted system, they are more impervious to tampering. The disadvantages include failure at wire speed (less suitable for high speed networks like 1G/10G). NIDS focuses on detecting the attacks from outside rather than the insider attacks. Further the deployment of NIDS can be
Distributed (Edge): NIDS can be deployed in the distributed manner in all the edges of the network. These types of NIDS have to communicate with each other.
Centralized: NIDS can be deployed centrally in the network where all the hosts get connected and in this type the load on the IDS will be high.
NIDS is generally installed the following strategic location of the network
<u>Before Gateway Firewall</u>*:* If the NIDS is installed before the gateway firewall, it can trace all the network events of interest. This has to handle all the incoming and outgoing traffic, this should be installed on the system with adequate resources such as processors, memory, network bandwidth etc. Configuring this type of deployment is very important and incorrect configuration may lead to lots of false alarms.
De-Militarized Zone (DMZ): Installing the firewall in-front of the DMZ, this enables the monitoring the traffic which are flowing to the public/internet-facing servers. Generally this is a second level deployment as this monitors the traffic already filtered by Gateway firewall.
Inside the Private/Corporate Network: Here the NIDS is deployed inside/within the corporate network. This will monitor the attacks emerging from inside as well as from outside. The scope of such deployments is very limited.
A Network IDS can be deployed in two modes
Inline Mode/Active: In inline mode the IDS is deployed in a strategic location where all traffic must pass through this IDS. In general, Inline IDSs are placed where the network firewalls and other security devices would be placed [25].
Out of Band/Passive Mode: In this mode the IDS monitors the copy of the actual network traffic and no traffic actually passes through the sensor. Passive sensors cannot stop any malicious packet instantaneously. Passive Sensors monitors the traffic by either Spanning Port or Network Tap.
Hybrid (Host-Network)**:** This combines the NIDS and HIDS approach and has the advantage of both the schemes.
Target Based: Target based IDS is a variation of standard host based IDS. In systems where resource based auditing is a constraint, this approach is used. Critical components of the network are identified and the IDS is placed at the entry/exit points.
Note: In the recent past, Intrusion detection research is mostly concentrated on anomaly based network intrusion detection [20].

C. INFORMATION SOURCE

IDS need data to process and take action. The information source for IDS can be either Audit trails or Network Packets.
Audit trails: This includes system/application logs, event logs etc. The host agents are deployed on each host to collect this information.
Network Packets: Every IDS includes a network based sensor to capture and process the network packets. Network packets are the main source of such IDSs.

D. IDS RESPONSE AND ANALYSIS FREQUENCY

Whenever an attack is detected, the IDS should respond in a predefined way.

The response may range from alert notification to block the network connection itself. The responses shall be based on the attack type and the security requirement of the organization. The responses can be of below types.

Passive Response: IDS can send alert messages to the administrator and the decision will be taken by the administrator. Alert messages and notifications can be sent on email, SMS and SNMP messages.

Actively defend: IDSs itself cannot block the attack but IDSs can block the network packets to avoid any potential damage. By sending TCP reset packets by reconfiguring firewalls or it can block the network traffic itself. This again can be of two types namely Proactive and Reactive.

"A proactive IDS instead will take pre-emptive countermeasures, like, actively interrogating all extant user processes and stopping all processes which did not originate from bona fide users at approved sites" [22].

Proactive IDS can be called as Intrusion Prevention System that has blocking capabilities depending based on the source of the packet and an IDS with Deep Packet Inspection (DPI).

IPS is the extension of ID with exercises of access control to protect computers from exploitation

Deep Packet Inspection (DPI) is a technology that enables the network owner to analyze internet traffic, through the network, in real-time and to differentiate them according to their payload (Content of data packets). IPS can be further classified into two types namely Host based Intrusion Prevention Systems (HIPS) and Network based Intrusion Prevention Systems (NIPS).

The present day IDSs rely on both the above mentioned response schemes and they are called as Intrusion Detection and Prevention Systems (IDPS).

E. DATA ANALYSIS FREQUENCY

The analysis of data captured by IDS can be analyzed in real-time or on batch mode later. The real-time analysis based IDSs process the data on the fly. With the present day network speed and bandwidth it is becoming difficult to process the captured data online. So organizations have started using number of sensors at various points/locations and for different purposes. In some cases the captured data needs to process in batch to correlate the activities/incidents. Generally audit trails are processed in batch and in forensics analysis this type of analysis will be helpful.

7. SNIFFING THE NETWORK

For monitoring or analyzing the network for any possible intrusions the network packets which pass thorough the network is to be made available to IDSs. The process of capturing network packets/flows is referred to Sniffing the network. There are various ways in which the network packets/flows can be eavesdropped without affecting the normal flow.

Packet capture by placing a Hub: Network hubs are physical layer device which broadcasts the entire packet it receives to all the ports and the destination devices processes packet and all the other devices connected to the Hub discards the packet. In this scheme IDS is connected to a Hub and the Network Interface Card (NIC) is configured in a promiscuous mode.

Port Mirroring or SPAN port: In a switched environment, the packets are forwarded only to the destination address as specified in the network packet unlike network hub where it is broadcasted to all the ports. Port mirroring or SPAN port is a technique in switched network where the designated mirrored port gets all the traffic. IDS is connected to the mirrored port in promiscuous mode so that it can process all the packets. The problem faced by port mirroring is that there are chances of packet drops because of traffic aggregation.

Network Tap: Network Tap is a hardware device which as three ports one for incoming, the second one for outgoing and the third port is a test port where all the incoming and outgoing traffic is seen. The IDS is conned connected to the test port to analyse all the packets. It does not introduce any delay since it does not process any packet.

Stealth Mode: IDSs should operate in stealth mode i.e. transparently since the attackers/intruders may try to target the IDS itself. The IDS port is set in promiscuous mode without assigning IP address and this arrangement only listens to the packets/flows keeping it transparent. IDS can be configured in stealth mode and packets can be sniffed by any of the above three modes.

Sniffing is the process of capturing, examining, and analyzing packets traversing the network. There are lots of tools available for packet sniffing. Some commonly used tools are discussed below.

Tcpdump: It is packet analyzer tool used by Information Security professionals and it enables the capture of packets which then can be saved and viewed. This is available in almost all flavors of UNIX.

Wireshark: It is the most commonly used told for network protocol analysis and it is free, open source tool. It is the de facto standard tool used by many especially government and educational institutions. It is used for network troubleshooting, analysis, and education. Originally named Ethereal, the project was renamed Wireshark in May 2006

NetowrkMiner: NetworkMiner is a Network Forensic Analysis Tool for Windows. NetworkMiner can be used as a passive network sniffer/packet capturing tool in order to detect operating systems, sessions, hostnames, open ports etc. without putting any traffic on the network. NetworkMiner can also parse pcap files for off-line analysis and to regenerate/reassemble transmitted files and certificates from pcap files

Netstumbler: Netstumbler is the best known Windows tool for finding open wireless access points. This is also available for WinCE version for PDAs called MiniStumbler. This is a free tool available on windows.

Net2pcap: This tool is used to capture network packet and convert it into a pcap file. This tool is generally used to capture packets and for subsequent analysis.

Snoop: This is a Linux tool which functions like tcpdump but uses a different format specified in RFC 1716. To observe the traffic between two systems this tool can be used.

Argus: This tool is available in most operating systems. It can process live data or captured data such as pcap files and the outputs the network flows.

8. ANOMALY DETECTION TECHNIQUES

Anomaly detection is the process of identifying anomalous events that occur in the system/ network with respect to the normal behavior with the basic assumption that attacks differ from the normal behavior. Normal events are to be identified and modelled. The main issues with these systems are the systems behavioral changes and the adaptability of anomaly detection system to these changes. Anomaly detection can be either supervised or unsupervised. Anomaly detection is a two-fold process where in the first phase the model is generated from a normal data and in the second phase using the model the possible deviations from the normal data is found out. The normal data from which the model is created should be a clean data without any noise, redundant information or missing values etc. The following are the major types of anomalies.

Point anomalies: If an individual data instance can be considered as anomalous with respect to rest of the data (e.g., if it lies outside the boundaries of normal region of the data).

Contextual anomalies: If an information occurrence is anomalous in a precise context, but not or else, then it is characterizing a related anomaly.

Collective anomalies: Collections of data instance are anomalous with respect to the entire data set; it is termed a collective anomaly.

Anomaly detection has been studied using various techniques and application domains. Anomaly detection methods can be broad classified into the following methods.

Statistical Methods: Statistical methods measure the system/network behavior over a period of time build model for normal behavior. The actual value is then compared with the model to find out anomalies. The basic assumption in statistical anomaly detection is the anomaly is irreverent because it is not generated by the scholastic model assumed [23]. Normal data falls in the high probability region and anomalies falls in the low probability regions. Examples include IDES, NIDES, EMERALD, PHAD, ALAD, LERAD and Haystack etc. This can be further classified into parametric and non-parametric methods. In Parametric methods, the knowledge of the distribution of the population is known and is in case of non-parametric methods the distribution is not known.

Parametric based: Parametric based methods assume that normal data is based on parametric/known distribution such as normal distribution. NIDS involves huge volume of high dimension data and intrusion to be detected early before intrusion causes damage. Therefore anomaly detection should use minimum computational cost. However in real-life the underlying distribution is not known. Chi-Square test is commonly used in Anomaly based Intrusion Detection.

Non Parametric based: Here the model is not defined priori. This methods does not assume any model initially but tailors the detection mechanism according to the data.

The following are some of statistical anomaly detection techniques

Operational Model (or) Threshold Metric: Operational Model (or) Threshold Metric: Anomaly is identified based on the comparing the observation with predefined limit. The cardinality of observations over a period of time is the basis for raising an alarm. The deviations in the count of certain values are often associated with anomalies/intrusions. The sudden surge in the SYN packets [26] or the number of password failure [27] for a brief period of time/interval may be an intrusion.

Markov Model: The normal behavior of an event is determined by using the observed characteristics of immediately preceding the events in Markovian model. This model characterized each observation as a specific state and utilizes a state transition matrix to determine if the probability of the event is high or normal, based on preceding events. It is useful when sequence of activities is particularly important [28]. This model is useful if the particular sequence of activities is important. In intrusion detection HMM is useful in maximizing detection rate and minimizing false positive error. This gives good performance at the cost of significant time required to model normal behaviors and determines intrusions. It is therefore not much suitable to real-time intrusions detection [29]. Markov chains and Hidden Markov models are two approaches in Markov Models.

Statistical Moments (or) Mean and Standard Deviation Model: In statistics mean, standard deviation or any other correlation is known as a moment. If any events moment is outside the specified confidence range then it is declared as anomalous. These models can learn the behaviour over a period of time and does not require any knowledge about the normal activity initially to set the limits. This model offers more flexibility than the threshold model [30].

Multivariate Models: This Model is similar to Statistical Moments model except that this is based on correlations amount two or more attributes. If two or more features are related to each other, this model will be useful Chi-Square [χ^2] test, t-test, Multivariate Cumulative Sum (CUMSUM) and Multivariate Exponentially Weighted Moving Average (MEWMA) are some of the examples of Multivariate Techniques [31]. As network traffic data has some features/attributes are related to each other, these models will be useful.

Time series Model: In time series model, the order and time interval of activities of the observation are reviewed to detect anomalies. If the probability of the occurrence of the event is low then it is termed as anomaly. Anomalies in time series are data points that deviate from the normal data patterns of the data sequence. Time series models are computationally expensive

Probabilistic Learning: The main characteristic of probabilistic learning is its ability to update previous outcome estimates by training them with newly available evidence. The independent works of Dempster and Shaffer on theory of belief and the original works of Bayes are basis for Bayesian Belief Networks. The following methods are some the techniques used in Probabilistic Learning.

- Hidden Markov Models (HMM),
- Bayesian network (BN),
- Naïve Bayes Technique,
- Gaussian Mixture Model (GMM),
- Expectation-maximization (EM)

Machine Learning (ML) and Data Mining (DM): These terms are used interchangeably and lot of confusion exists about these terms. Data Mining is the process of extracting implicit, previously unknown and potentially useful information from the data. Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term "machine learning" in 1959 while at IBM. “ML is the field of study that gives computers the ability to learn without being explicitly programmed". ML provides algorithms that resolve the problem based on the data, and the solution that improves with time. Data Mining is used to extract regularities from a very large database as part of a business cycle. DM focuses on the discovery of previously unknown properties in the data.

ML focuses on classification and prediction, based on known properties previously learned from the training data. An ML approach usually consists of two phases: training and testing. The detection models are constructed based on the past behaviour. The learning algorithm analyses the data e.g. network packets/flow and build models on the normal behaviour. Machine learning based detector then uses the model to detect the deviations from the normal pattern. These algorithms can be modelled to adapt itself to the changes. The following are some of the algorithms which are supervised.

Decision Trees: Decision trees are structures used to classify data. Decision trees represent a set of rules which categorize data based on the values of the attributes. Decision trees are popular tool for classification and prediction. A decision tree is a tree that has three main components: nodes, edge, and leaves. A decision tree can be used to classify a data point by starting at the root of the tree and moving through it until a leaf node is reached. The leaf node would then provide the classification of the data point. There are two methods of building a tree top-down and bottom-up. ID3 and C4.5 are commonly used algorithms of decision trees which uses top-down approach. The decision tree has to be first trained with known data before it can be used to classify unknown data.

Support Vector Machines (SVM): SVM is a supervised machine learning algorithm used for both classification and regression analysis. The basic learning process includes mapping the original input space to a higher dimensional (n-dimensional) feature space and finalizing a hyper plane within the feature space with a minimum margin using Sequential Minimum Optimization (SMO) or other methods.

Bayesian Networks: Bayesian network or a belief network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies in a form of directed acyclic graph [31]. Each node corresponds to a random variable and edges represent relations. The nodes that are not connected represent variables that are conditionally independent from each other. The Bayesian Network learns the casual relations between attributes and class labels from the training dataset before it can classify unknown data.

K-Nearest Neighbor: The k-nearest neighbors' algorithm (k-NN) is a non-parametric method used for classification and regression.

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k=1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

The following are some of the algorithms which are unsupervised

K-Means Algorithm: The k-means algorithm [32] is one of the most well-known centroid algorithms. It partitions the dataset into k subsets such that all points in a given subset are close to the same centre. It randomly selects k of the instances to represent the cluster centres and based on the selected instances, all remaining instances are assigned to their nearest cluster centre. K-means then computes the new cluster centers by taking the mean of all data points belonging to the same cluster [20]. The process is iterated until some convergence criterion is met. The four important properties of k-means algorithm are

(i) it is scalable, i.e., can handle large datasets,
(ii) it often converges to a local optimum,
(iii) it can identify clusters of spherical or convex shapes
(iv) It is sensitive to noise.

A major limitation of this algorithm is that the number of clusters must be provided as an input parameter. In addition, choosing a proper set of initial centroids is a key step of the basic k-means procedure and results are dependent on it.

Expectation Maximization (EM): EM algorithm is a general method for finding maximum likelihood estimation (MLE).of the parameters of a distribution from a given dataset, even if the data is incomplete or has missing values. The EM algorithm starts with an initial guess for the parameters of the mixture model for each cluster and then iteratively applies the Expectation Step (E) and the Maximization Step (M), in order to converge to the maximum likelihood fit. In the E-step, the EM algorithm first finds the expected Value of the complete-data log-likelihood, given the observed data and the parameter estimates. In the M-step, the EM algorithm maximizes the expectation computed in the E-step. The two steps of the EM algorithm are repeated until an increase in the log-likelihood of the data, given the current model, is less than the accuracy threshold. EM algorithm is used to cluster incoming audit network data and compute missing values [33].

Machine Learning approaches that have been developed are new and still evolving. From the above literature review the need for a hybrid model with the combination of classification approaches are required to make the decision intelligently for improving the overall performance of the IDS.

Data mining based anomaly detection discovers consistent patterns from the behaviour of network traffic. Data mining combines algorithms used in different methods like machine learning and signal processing etc. [50]. Researchers are looking at data mining technique for the elimination of manual and ad-hoc processes for building Anomaly Detection Systems (ADS). Data mining methods excels at processing large volume of data. The data mining process is likely to reduce the data being processed for historical comparison of network activity, creating more meaningful data [35, 20]. Data Mining can be further classified into

Clustering: This is an unsupervised technique for discovery of pattern in unlabelled high dimension data. Clustering is the process of making a group of abstract objects into classes of similar objects. Cluster is a group of objects that belongs to the same class. If the test record is at a longer distance from the normal cluster then the same may be classified as attack. The choice of clustering depends on the type and dimension of data. Clustering methods can be classified into the following categories.

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Classification: It is the process of finding a model (function) that describes/distinguishes data classes for the purpose of being able to use the model to predict the class of objects whose class is unknown. The model is based on the analysis of training data whose class is known.

Association: Association discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rules are if/then statements that help uncover relationships between seemingly unrelated data. Given a set of transactions, Association is the process of finding, rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Outlier Mining: The aim of Outlier mining is to identify patterns in data that do not conform to the rest of the data based on approximate proximity measure. The non-conforming patterns are referred to as outliers, noise, exceptions, anomalies, errors, etc. The following are some the outlier mining approaches.

Distance based outlier detection
Density based outlier detection
Outlier detection based on soft computing

Soft Computing approaches: Soft computing techniques employ approximate reasoning with necessary tolerances to mimic human mind. The following are some of the soft computing techniques. It is an innovative approach to build a computationally intelligent system, analogous to the reasoning of human mind and ability to learn from environment under imprecision and uncertainty [34].

Artificial Neural Network (ANN): It is an information processing model that is inspired by the way the information is processed in biological nervous systems such as brain. ANN consists of lots of processing elements called neurons working in unison to solve problems. ANN is also called as Neural Networks (NN). The main advantage of Neural Networks is that it can learn and adapt the behaviour to the changing condition [36]. ANN is very useful in detecting anomalies in network traffic.

Fuzzy Logic: The concept of Fuzzy Logic (FL) was conceived by Lotfi Zadeh, a Professor at the University of California at Berkeley. This method is used in situations where imprecise data needs to be processed. The data are considered as fuzzy sets. Fuzzy logic is a form of multi-valued logic derived from fuzzy set theory to deal with reasoning that is approximate rather than accurate. Fuzzy systems remains controversial among statisticians but this has achieved commercial applications.

Genetic Algorithm: Genetic Algorithms deals with evolutionary processes based on the concept ‘Survival of the fittest’. It was originally introduced in the field of computational biology. The selection and evolution process is implemented in computers using Genetic Algorithms (GA). Genetic Algorithms eliminate redundancy. GA’s are widely used as searching algorithms. GA reduces the feature subsets while maintaining or improving learning accuracy is maintained. Genetic algorithms [37] are also effective in cluster analysis. GA-based clustering uses randomized search and optimization based on the principles of evolution and natural genetics.

Rough Set: Rough set is a new mathematical theory for dealing with Vagueness and uncertainty. Machine learning-based intrusion detection approaches have been subjected to extensive researches because they can detect both misuse and anomaly. Rough set classification (RSC), a modern learning algorithm, is used to rank the features extracted for detecting intrusions and generate intrusion detection models [22].

There are other methods such as ACO, PSO, Evolutionary Computing and Reinforced Learning etc. which are beyond the scope of this thesis.

Knowledge Based: In this method, the events on the network/host are checked against the predefined rules/patterns of an anomaly. Examples include Rule based, Ontology based, Logic based, State Transition based, Petri Nets based and Expert System based systems.

Spectral Anomaly: Spectral techniques try to find an approximation of the data using a combination of attributes that capture the bulk of the variation in the data. Data is projected on a lower dimensional subspace in which the normal and attack instances appear significantly different. Techniques such as Principal Component Analysis (PCA) are used to project the data in lower dimensional space.

Protocol based anomaly detection: These systems monitor the protocol for any deviation in the standard specification. Most of the protocol based anomaly detectors are built as state machines i.e. the detector is monitoring the transition from one state to another and if the anticipated transition is different from the transition that has occurred then it is defined as anomaly [38].

In this section various anomaly detection methodologies for Anomaly detection has been discussed under different categories. Each of these techniques has its own advantages and disadvantages and the selection of the methods has be done based on the need. Most of these techniques assume that anomalies are rare compared to normal data. The following part of the thesis do assume the same way and in any network if the anomaly is more than 10% then it calls for a serious concern.

9. ANOMALY DETECTION SYSTEMS

In this section different anomaly detection systems are discussed based on the core functionality.

Application Layer Anomaly Detector (ALAD): ALAD [39] is designed to detect the attacks at application layer. A model on normal behaviour is built and this is used to detect anomalies. Anomalies are detected in inbound TCP connection to known ports. This falls under the category of Statistical based anomaly detector [40]. It uses a time-bound model. ALAD was trained and tested using DARPA dataset. 70 known attacks were detected out of 180 and the false alarm was 100.

Packet Header Anomaly Detector (PHAD): The normal range of values Packet headers of Ethernet, IP, TCP, UDP, ICMP learned and connections are examined for possible intrusion. Attribute scores are calculated and they are aggregated to for packet anomaly score. The reliability of PHAD is higher in detecting Probe attacks.

Learning Rule for Anomaly Detection (LERAD): This uses a rule based learning algorithm to pick up good rules rather than fixed set of rules and it is first method to characterize the normal behaviour in the absence of normal data [41]. It discovers the relationship between attribute within a packet.

Minnesota Intrusion Detection System (MINDS): This is one of the popular systems based on data mining. MINDS use Clustering and Outlier Detection to detect density-based local outliers that indicate intrusions. It uses net flows as the source and uses an outlier algorithm to assign anomaly scores to each connection. Output of MINDS anomaly detector consists of original Net flow data along with the anomaly score and relative contribution of 16 attributes used by anomaly detection algorithm [42].

Automated Data Analysis and Mining (ADAM): This system uses data-mining based rule association and classification techniques for identifying incursions. ADAM uses tcpdump as the information source. Filtering technique is used initially to filter out normal events and classification is performed to identify attacks [43]. ADAM both attack-free traffic and traffic with labelled attacks.

Mining Audit Data for Automated Models for Intrusion Detection (MADAM ID) uses data mining to audit the captured behavioral patterns of intrusion and normal data. The performance of MADAM ID was good in detecting known attacks in 1998 DARAPA evaluation dataset [44]. This project at Columbia University has shown how data mining techniques can be used to build IDS. MADAM ID has been documented for network misuse detection

Graph-based Intrusion Detection System (GrIDS): GrIDS was developed during 1996, which codes the system/hosts as nodes and connection between them as edges. These graph present network events in a graphic fashion that enables the viewer to determine if suspicious network activity is taking place [45]. Graph engines are built hierarchically. Higher level engines gather data from lower level graph engines and at the lowest level from single hosts. Lower level graphs pass information to higher level engines, and graphs become coarser and coarser on each upper level.

Rate-Limiting: The underlying assumption of this system is that the affected host will transmit more data or try to connect too many systems in a short period of time. Therefore the traffic which initiates more connections at a higher rate is delayed. Generally this work on the server side and monitors the outgoing connection [46].

Next-generation Intrusion Detection Expert System (NIDES): This system is based on statistical approach. It is a distributed IDS and is a combination of both signature based and anomaly based detection model [12]. NIDES operate in real-time as well as batch mode.

PAYL: This system is based on payload and the main assumption is that attacks are very less when compared with the normal packets and can be readily identified. This system can be trained even if there is traces of noise are present. The major disadvantage of this method is complexity and while processing the payload it fails to detect attacks related to packet headers [47]. This Models the normal application payload of network traffic in a fully automatic, unsupervised and very efficient fashion. This system Computes profile byte frequency distribution and the standard deviation of the application payload flowing to a single host and port during the training phase.

Fuzzy Intrusion Recognition Engine (FIRE): This is an anomaly based IDS that uses fuzzy logic. It generates fuzzy sets for every observed feature which are in turn used to define fuzzy rules to detect individual attacks [15].

NETAD: NETAD like PHAD detects anomalies in network packets. PHAD takes into account only the first 48 bytes of each packet whereas NETAD use those 48 byte as one of the attributes. NETAD worked well with DARPA data on 18nattacks which are poorly detected by other IDS [48].

Event Monitoring Enabling Responses to Anomalous Live Disturbances (EMERALD): It is a hierarchical intrusion detection system that monitors the hosts, domains, etc. EMERALD employs ensemble of techniques like statistical analysis and expert system. It is a hybrid of signature based and anomaly based system.

10. MORPHOLOGICAL ANALYSIS OF IDS

“Morphological analysis is simply an ordered way of looking at things” [49]. Morphological Analysis is a simple, powerful conceptual methodology widely used in linguistics, biology and technology forecasting. Today, morphology is associated with a number of scientific disciplines in which formal structure is a central issue. However, its use can easily be extended to many other areas of knowledge.

Morphological frameworks are characterized by dimensions and options for each dimension. Dimensions refer to component parts of the structure of an entity under study. The analysis phase begins by identifying and defining the most important dimensions of the problem complex to be investigated. Each of these dimensions is then given a range of relevant values or conditions. Together, these make up the variables or parameters of the problem to be structured. A morphological field is constructed by setting the parameters against each other, in parallel columns, representing an n-dimensional configuration space. The morphological analysis enables the identification of new dimensions and options and becomes a powerful tool for conceptualizing innovations. A morphological framework of IDS is presented in Table 3.

Table 3: Morphological Analysis of IDS

IDS Dimensions	**Options**
Detection Methods	Signature/Pattern based/Misuse/Knowledge based, and Behaviour based /Anomaly based.

Learning Techniques	Supervised, Un-supervised and Semi-Supervised
IDS positioning	Host-based, Network based (distributed/edge, centralized), Hybrid, Target based
Installation location	Before gateway firewall, De-Militarized Zone(DMZ), Inside Private/Corporate Network
Modes of Deployment	Inline/Active Mode, Out of Band/Passive Mode(Spanning port, network tap)
Information Source	Network Packets and Audit trails
IDS Response	Active defend and passive response
Data Analysis Frequency	Real-time and batch mode
Packet sniffing methods	Packet capture by hub, Port mirroring or span port and network tap, Stealth mode
Type of anomalies	Point anomaly, context anomaly and collective anomaly
Anomaly detection Methods	Statistical, Machine Learning, Probabilistic Learning, Data Mining, Soft Computing, Knowledge or Rule Based, Spectral Anomaly, Protocol based anomaly
Anomaly Detection Systems	ALAD, PHAD, LEARD, MINDS, ADAM, MADAM, GrIDS, NIDES, Rate Limiting, PAYL, FIRE, NETAD, EMERALD,
Type of datasets used in evaluation of IDS	Real Traces, Benchmark, Synthetic
Metrics for evaluation of IDS	Detection Rate(Precision), False Alarm Rate, Accuracy, Sensitivity(Recall),F-score, MCC

The above Morphological framework is used for understanding various dimension and options for IDS.

11. SUMMARY

This paper outlines the concepts, methodologies and literature surrounding IDS are discussed in detail and this form the Body of Knowledge (BoK) of IDS. The Morphological analysis and taxonomy of IDS are discussed in detail.

12. REFERENCES

1. Anderson, J.P., Computer Security Threat Monitoring and Surveillance, Technical report, James P. Anderson Co., Fort Washington, PA., April 1980. On Software Engineering, vol. SE-13, pp. 222-232, February 1987.
2. Ashok Kumar, D., and Venugopalan, S.R., 2016, December. A Novel algorithm for Network Anomaly Detection using Adaptive Machine Learning. In Advanced Computing and Intelligent Technologies (ICACIE), 2016 First International Conference on. Springer
3. Singh, S.P. (2010) Data Clustering Using K-Mean Algorithm For Network Intrusion Detection, Thesis, Lovely Professional University, Jalandhar.
4. Deepthy K. Denatious, and John, A. (2012) ‘Survey on data mining techniques to enhance intrusion detection’, International Conference on Computer Communication and Informatics, ICCI-2012, Coimbatore, India.
5. C. Kruegel, F. Valeur, and G. Vigna. Intrusion Detection and Correlation: Challenges and Solutions. Springer-Verlag Telos, 2004.
6. L. R. Halme and R. K. Bauer. AINT misbehaving – A taxonomy of anti-intrusion techniques. In Proc. of 18th NIST-NCSC National Information Systems Security Conference, pages 163–172, 1995.
7. D.E. Denning, An Intrusion-Detection Model, IEEE Transactions on Software Engineering, vol. SE-13, pp. 222-232, 1987.
8. Dinakara K, “Anomaly Based Network Intrusion Detection System”, Thesis Report, Dept. of Computer Science and Engineering, IIT Khargpur 2008
9. Guy Bruneau – GSEC Version 1.2f,” The History and Evolution of Intrusion Detection”, SANS Institute 2001.
10. Ilgun, Koral, USTAT:a real time IDS for Unix, Proceedings of the 1993 IEEE Computer Society Symposium on research insecurity and privacy, 1993.
11. Mark Crosbie, Gene Spafford, Defending a Computer System using Autonomous Agents, Technical report No. 95-022, COAST Laboratory, Department of Computer Sciences, Purdue University, March 1994.
12. D. Anderson, T. Frivold, A. Valdes, Next-generation intrusion detection expert system (NIDES), Technical report, SRI-CSL-95-07, SRI International, Computer Science Lab, May 1995."
13. Paxson, Vern, Bro: A system for detecting network intruders in real-time, Computer Network, v 31, n 23, Dec 1999.
14. Ning,Wang X.S, Jajodia S, Modelling requests among cooperating IDSs, Computer Communications, v 23, n 17, Nov, 2000."
15. J. E. Dickerson and J. A. Dickerson, “Fuzzy network profiling for intrusion detection,” In Proceedings of the 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS), 13-15 July 2000, pp. 301 – 306.
16. Debar H, Becker M, and Siboni D, “A Neural Network Component for an Intrusion Detection System”, IEEE Computer Society Symposium on Research in Security and Privacy, Los Alamitos Oakland, CA, pp. 240–250, May 1992.
17. Ghosh A, K. A Schwartzbard, and M Schatz, “Learning program behavior profiles
18. D. Barbara, N. Wu, and S. Jajodia, “Detecting novel network intrusions using bayes estimators”, In proceedings of the first SIAM international conference on Data Mining, Chicago , USA, Apr 2001.
19. Jiong Zhang and Mohammed Zulkernine, “Anomaly based Network Intrusion Detection with Unsupervised Outlier Detection”, IEEE International Conference on Communications 2006.
20. DK Bhattacharyya and JK Kalita, 2014, “Network Anomaly Detection: A Machine Learning Perspective”, CRC Press, Taylor & Francis Group, International Standard Book Number-13: 978-1-4665-8209-5
21. Bhuyan, M. H., Bhattacharyya, D. K., and Kalita, J. K. Surveying port scans and their detection methodologies. The Computer Journal 54, 4 (April 2011), 1-17.
22. Thomas, C., 2009. Performance enhancement of intrusion detection systems using advances in sensor fusion. Supercomputer Education and Research Centre Indian Institute of Science, Doctoral Thesis. Available at: http://www. serc. iisc. ernet. In/graduation-theses/CizaThomas-PhD-Thesis.pdf.
23. V. Chandola, A. Banerjee and V. Kumar. ACM Computing Surveys, Vol. 41(3) Article 15 2009. DOI 10.1145/1541880.1541882 http://doi.acm.org/10.1145/1541880.1541882.
24. Wikimedia, Foundation. Intrusion detection system. http://en.wikipedia.org/wiki/Intrusion-detection system, February 2009.

25. Longe Olumide Babatope., Lawal, Babatunde. Ibitola Ayobami, “Strategic Sensor Placement for Intrusion Detection in Network-Based IDS” I.J. Intelligent Systems and Applications, 2014, 02, 61-68, I.J. Intelligent Systems and Applications, 2014, 02, 61-68

26. Vasilios S.; Fotini P., “Application of anomaly detection algorithms for detecting SYN flooding attacks”, Elsevier, Computer Communications, Vol. 29, pp. 1433, 1442, 2006

27. Dorothy D., “An Intrusion-Detection Model”, IEEE Transactions on Software Engineering, Vol. SE-13, No. 2, pp. 222, 232, Feb. 1987

28. James C.; Jay H., “A Comparative Analysis of Current Intrusion Detection Technologies”, Proceeding of 4th Technology for Information Security Conference, TISC’96, Houston, TX, May.1996"

29. Anurag Jain, Bhupendra Verma and J. L. Rana., “Anomaly Intrusion Detection Techniques: A Brief Review”, International Journal of Scientific & Engineering Research, Vol 5(7), 2014

30. Manasi Gyanchandani, J. L. Rana, R .N. Yadav, “Taxonomy of Anomaly Based Intrusion Detection System: A Review”, International Journal of Scientific and Research Publications, Vol 2(12), 2012

31. Martin Elich, “Flow-based Network Anomaly Detection in the context of IPv6”, Thesis Report, FAKULTA INFORMATIKY, MASARYKOVA UNIVERZITA, 2012.

32. Hartigan, J. A., and Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. Applied Statistics 28, 1 (1979), 100-108.

33. Patcha, A., and Park, J.-M. Detecting denial-of-service attacks with incomplete audit data. In Proc. of the 14th Int'nl Conference on Computer Communications and Networks (ICCCN 2005) (October 2005), IEEE Computer Society, pp. 263-268."

34. Sampada Chavan, Khusbu Shah, Neha Dave and Sanghamitra Mukherjee” Adaptive Neuro-Fuzzy Intrusion Detection Systems” Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC’04) IEEE 2004.

35. Narayana; Prasad; Srividhya; Reddy, “Data Mining Machine Learning Techniques – A Study on Abnormal Anomaly Detection System”, International Journal of Computer Science and Telecommunications, Vol. 2, Issue 6, Sept. 2011

36. Yevgeniy Bodyanskiy, Sergiy Popov, Neural Network Approach to Forecasting of Quasiperiodic Financial Time Series, European Journal of Operational Research Vol. 175, pp. 1357-1366, 2006.

37. Goldberg, D. E. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, New York, 1989.

38. Das, K. Protocol Anomaly Detection for Network-based Intrusion Detection, SANS Institute, GSEC Practical Assignment Version 1.2f, 2001

39. M. V. Mahoney and P. K. Chan, “Learning Non stationary Models of Normal Network Traffic for Detecting Novel Attacks.” ACM SIGKDD international conference on Knowledge discovery and data mining, 2002.

40. ACM Press, “Learning non stationary models of normal network traffic for detecting novel attacks,” in Eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 2002, pp. 376–385.

41. Chan, P. K., Mahoney, M. V., and Arshad, M. H. A machine learning approach to anomaly detection. Tech. Rep. CS-2003-06, Department of Computer Science, Florida Institute of Technology, 2003

42. Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Kumar,V., and Srivastava, J. MINDS | Minnesota Intrusion Detection System, 2004.

43. D. Barbar´a, J. Couto, S. Jajodia, and N. Wu, “ADAM: a testbed for exploring the use of data mining in intrusion detection,” in ACM SIGMOD Record: SPECIAL ISSUE: Special section on data mining for intrusion detection and threat analysis, vol. 30, no. 4. ACM Press, 2001, pp. 15–24.

44. Lippmann, R. P., Fried, D. J., Graf, I., Haines_ J. W., Kendall, K. R., Mc-Clung, D., Weber, D., Webster, S., E., Wyschogrod, D., Cunningham, R. K., and Zissman, M. A., (2000)

45. S. Staniford-Chen, S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, C. Wee, R. Yip, D. Zerkle, GrIDS – A Graph-Based Intrusion Detection System for Large Networks, The 19th National Information Systems Security Conference, Baltimore, MD., October 1996.

46. M. M. Williamson, “Throttling viruses: Restricting propagation to defeat malicious mobile code,"" ACSAC Security Conference, 2002.

47. K. Wang, S. Stolfo, “Anomalous Payload-Based Network Intrusion Detection,” Recent Advances in Intrusion Detection (RAID), 2004.

48. M. Mahoney, “Network Traffic Anomaly Detection Based on Packet Bytes,” ACM Symposium on Applied Computing (SAC), 2003.

49. Zwicky, F. (1948a). Morphological astronomy. *Observatory*, *68*(845), 121–143.

50. Lee, W., Stolfo, S. J. Data Mining Approaches for Intrusion Detection, Proceedings of the 7th USENIX Security Symposium, pp. 26-29, San Antonio, Texas, January 1998.

51. Martin Roesch: “Snort Documents”, http://www.snort.org/docs/ 1998